TS-EAS Shared Schema Subteam - Report (July 2019)

Submitted by: Kerstin Arnold (chair)
Members: Karin Bredenberg, Katherine Wisser, Kerstin Arnold, Mark Custer / Erica Boudreau, Regine Heberlein

Repository: https://github.com/SAA-SDT/SharedSchema

Timeframe: January to July 2019, (near to weekly) online meetings (approx. 20 meetings, 1 hour each)

Coverage at the time of reporting

This report on shared schema considerations refers to the current EAD3 (Version 1.1.0) and EAC-CPF (Edition 2018). Basis for the conversations within the subteam has been a comparison between these two, which was compiled by team members Erica Boudreau and Katherine Wisser during 2018 and updated by the subteam during 2019.

Extended coverage for next steps

Given the ongoing work of TS-EAS with regard to introducing a new standard on functions (EAC-F) as well as the TS-EAS endorsed work led by the Working Group on Standards of the Archives Portal Europe Foundation on EAG, these two formats should ideally be included in future considerations, provided this piece of work is taken forward. Specifically with regard to EAC-F it would then also be anticipated, that the work on a shared schema will inform the development of this standard accordingly.

Work of the Shared Schema Subteam so far

The Shared Schema Subteam was initiated in January 2019 with the task to:

- examine the idea of a shared schema and looking at pros and cons (in general);
- identify specific areas that could be included in a shared schema for EAD3 and EAC-CPF (future revised edition), keeping the development of a potential EAF in mind;
- make recommendations based on the above for TS-EAS to consider.

The subteam then met weekly (as far as possible) to discuss the various aspects of their task in detail.

Based on the pre-existing comparison of EAD3 and EAC-CPF, the subteam looked at:

- elements and attributes that are currently shared between both standards;
- elements and attributes that are currently specific to the one or the other standard.

The main criteria for distributing elements and attributes in one of these two groups were their names, which means:

- the element names as e.g. used in the tab libraries;
- the tag names as they would show in the XML files

With regard to the tag names, the aspect of camel-casing was set aside for the time being. I.e. elements like <bioghist> (EAD3) and <biogHist> (EAC-CPF) were considered as the same element.

Looking at the elements and attributes of both groups in more detail - and with a joint approach rather than from the specific perspective of the one or the other standard - quickly showed that the initial distribution was not necessarily clear-cut in each case:

- elements and attributes with the same name in both standards might have a very different scope and/or purpose in the one or the other;
- elements and attributes with different names in both standards might actually have a very similar, if not the same scope and/or purpose.

Hence, the group decided on the following general principles rather early on, which should apply independent from whether or not there would be a shared schema in future:

- If elements and attributes in multiple standards are “meant to do the same thing”, they should be named the same and should do the thing they do in the same way.
- If elements and attributes in multiple standards are “meant to do something different” from each other, they should be clearly distinct from each other (i.e. have different names, use different default values, emphasise differences in their descriptions and the accompanying examples in the tag library etc.).
- Such similarities - or differences - should not only be reflected in the tag names and the schema definitions, but as well in the element names, the descriptions and examples as provided in the tag libraries.
- Ideally, the combination of the above will lead to minimising the number of (different) ways to do the same thing by:
    - Being as consistent as possible across all standards;
    - Reducing ambiguity as far as possible;
    - Making scope and/or purpose as explicit as possible.

Harmonisation and modularisation

EAD and EAC-CPF, so far, have been developed by different groups of people at different times. This has resulted not only in “personal preferences” shining through, but also in different stages of standards development more generally being reflected in the current variations of encoding possibilities. As TS-EAS has now been reorganised into one group for both standards, and as both standards are used alongside each other more and more, the question of harmonisation finds itself to be more in the centre of attention. Aiming to

harmonise EAD and EAC-CPF as far as possible and keeping them only as different as they absolutely need to be, could be a big step towards making it easier to maintain, to teach, to adopt and to implement those standards.

Using a modular approach along with schema annotations, we have the option to define shared elements in different ways within our different schema deliverables for each standard. For just one example, even if we define the <control> element once, we could allow EAD's profile of the control element to include a <filedesc> element, whereas the EAC version would continue to exclude filedesc. Alternatively, if we wanted a fully-harmonised <control> section, then we could either opt to promote <filedesc> to become a sibling element of <control> in EAD3 2.0 or consider adding <filedesc> as sub-element of <control> in the next version of EAC-CPF.

Since modularising the schemas will make the schemas extensible, it will be up to us to decide how similar (or not) we would like to keep each of the modules. Therefore, whether we have any interest whatsoever to e.g. allow an attribute like @audience or @encodinganalog in EAC or not, or whether we want EAD to benefit from improvements that are made during EAC's major revision or not, either decision is possible to accommodate.

Technical aspects of a shared schema approach

Should TS-EAS decide to move forward with the idea of a modularised shared schema, there are various more general questions to be considered as part of this approach:

- Namespace;
  EAD (for both its major versions) and EAC-CPF currently have separate namespaces. In a first step of harmonisation, this could be kept as is, while considering to move the EAC-CPF namespace from its current "urn" format to an "http" format, which would remove the step of de-referencing that is required with a URN. Potentially in the context of changing EAC-CPF's namespace as part of the major revision, the next step of harmonisation could then be to move to a single EAS namespace for whatever schema(s) that TS-EAS provides.
- Naming conventions;
  Connected to the question of separate vs. joint namespace(s), there will be the question of harmonising the naming conventions - or not. Deciding for a single EAS namespace would come with deciding for a common naming convention (camelCase, PascalCase, lowercase, snake_case, etc.) for all TS-EAS schemas to avoid having cases such as eas:bioghist and eas:biogHist right next to each other. Keeping separate namespaces, i.e. having ead:bioghist and eac:biogHist, would still allow TS-EAS to decide for one joint naming convention when defining all elements and attributes (e.g. camelCase), while the schema publication process could take care of adapting that naming convention for the actual schema deliverables (e.g. lowercasing everything for EAD).
- Schema serialisation;
  Seeing the approach taken by the majority of other standards communities, TS-EAS should opt for providing the same schema serialisations for each schema, and only one serialisation for each schema. Having six different schema serialisations, plus a Schematron file, for EAD3 and three schema serialisations for EAC-CPF with - to

some extent significant differences - makes it not only difficult to maintain, but also difficult for users (and other potential implementers) to navigate and understand.

See supporting document on Shared Schema Considerations for more details.

Encoding aspects of a shared schema approach

To get an understanding of the extent of harmonisation that might be required if TS-EAS were to follow a shared schema approach, the subteam evaluated the information gathered in the comparison of EAD3 and EAC-CPF in more detail.

There is a total of 52 elements that are currently shared between both standards, which currently amounts to about a third of all EAD3 elements and more than half of all EAC-CPF elements. Furthermore, while using the same name:

- All of these elements differ with regard to the attributes that they are used with.
- 20 of these elements differ with regard to their content model, i.e. the sub-elements that they are used with.
- 10 of these elements differ with regard to being repeatable or not.
- 7 of these elements differ with regard to being mandatory or optional.
- 7 of these elements differ with regard to their purpose and scope.

On the other hand, there are 9 EAC-CPF elements the purpose and scope of which puts them in close relation to differently named EAD3 elements; and there are 16 EAD3 elements the purpose and scope of which suggests a certain proximity to differently named EAC-CPF elements.

This includes cases, where an element specific to one standard could be considered the equivalent or variation:

- of a group of elements specific to the other standard, e.g. <nameEntry> in EAC-CPF and the group of <corpname>, <famname>, <persname> and <name> elements in EAD3;
- of a shared element in both standards, e.g. <outline> in EAC-CPF and <list> in EAC-CPF as well as in EAD3;
- of a more generic element in the other standard, which might be too generic for the similarity to be applicable in both directions, e.g. <didnote> in EAD3 and <descriptiveNote> in EAC-CPF.

See supporting document on Similar Elements Analysis for more details.

Shared elements and attributes

Skipping the differences of shared elements with regard to being mandatory, optional or repeatable, there are three main aspects with according sub-questions to consider:

1. Differences in attributes used;
    a. Evaluate attributes that do not exist in the one standard and the other and decide whether these should be added. This would mainly lead to additions of attributes in EAC-CPF, if approved. Examples of “missing” attributes are @audience, @encodinganalog, or the group of @calendar, @era, and @certainty.

    b. Evaluate attributes that are used inconsistently, i.e. exist in both standards, but are “missing” in certain contexts of the one or the other. Cases like this can be found in both, EAD and EAC-CPF.
2. Differences in content model;
    a. Evaluate elements that do not exist in the one standard or the other and decide whether these should be added. Contrary to the same question for “missing” attributes, the question here also touches on the aspect of differences in scope and purpose (see point 3 below). E.g. <legalstatus> in EAD is a general descriptive element similar to the likes of <scopecontent> or <bioghist> and uses the m.blocks element group, while in EAC-CPF it is considered rather to be part of authority file terms such as <function> or <occupation> mainly allowing for the indication of relevant dates and places.
    b. Evaluate elements that are used inconsistently, i.e. exist in both standards, but are “missing” in certain contexts of the one or the other. This again is related to a difference in scope and purpose, so that some of these cases might be easier to “solve” than others, e.g. <dateSet> not being part of <chronItem> in EAC-CPF, while there is <dateset> in <chronitem> of EAD3, or <abstract> being used with <biogHist> in EAC-CPF, but not in EAD3, where <abstract> is very much bound to <did> and can summarise information from a variety of <did>’s sibling elements.
3. Differences in scope and purpose;
    a. Evaluate elements that are named the same, but appear to have a different logical focus based on description, examples and - ideally - real-life usage.
    b. Identify the subgroup of elements that are meant to be the same and the subgroup of elements that are meant to be different in order to decide on next steps (harmonisation or renaming and redefining respectively).

See supporting document on Shared Elements Analysis for more details.

Alternatives to a shared schema approach

Based on the general principles as identified in the chapter describing the work of the Shared Schema Subteam, the minimum action to take would be to:

- identify those currently shared and/or similar elements and attributes that are meant to be and do the same;
    - harmonise the names of these elements and attributes across both schemas in case they are not using the same names already;
- identify those currently shared and/or similar elements and attributes that are meant to be and do something different;
    - consider renaming these elements and attributes within their respective standards in case they are currently using the same names;
- make sure that the descriptions of elements and attributes in both tag libraries along with the examples provided show and explain especially the differences of seemingly similar elements and attributes.

Questions to TS-EAS and recommendations for next steps

The minimum action as detailed in the previous chapter of this report would already put TS-EAS in a position to:

- harmonise the EAS standards, which was one of the reasons of creating a single TS-EAS;
- refactor and fix bugs that exist in one or both of the separate schemas.

Deciding for a shared schema approach, however, would additionally give TS-EAS opportunity to:

- reduce the duplication of effort that is currently required to update both schemas (e.g. the recent example of adding the new “Rights declaration” element to both, EAD3 and EAC-CPF), which becomes even more relevant when considering future additions to the EAS suite such as EAC-F;
- enable implementers of the EAS suite to reuse rather than duplicate code.

The Shared Schema Subteam therefore kindly asks TS-EAS to:

- agree to following a shared schema approach and task the Schema Team with this work;
    - task the Schema Team with developing a set of first principles to guide schema design;
    - task the Schema Team with evaluating the impact of this new approach to schema design on the maintenance of the tag libraries;
    - task the Schema Team with investigating if there are any examples from the XML community where initially separate schemas have been merged;
- agree on a principle of naming conventions and namespaces;
- agree on a principle of schema serialisation;
- agree on a timeline for moving towards a shared schema,
    - which aligns with the timeline for the current major revision of EAC-CPF,
    - which allows a step-by-step integration into EAD3 as part of its rolling minor revision cycles, and
    - which takes next steps for further development of EAC-F into account.